BU-1097-MA

Assessing Influence in Variable Selection Problems

Christian Léger *
Département IRO
Université de Montréal

Naomi Simone Altman †
Biometrics Unit
Cornell University

February 1991

Abstract

Variable selection techniques are often used in combination with multiple linear regression to produce a parsimonious model that fits the data well. It is clearly undesirable for the final model to depend strongly on the inclusion of a few influential cases (data points) in the data set. This article discusses a measure of influence of single cases on the final model, based on a similar measure used in ordinary multiple regression.

When variables are selected objectively using the data, deletion of individual cases can strongly affect the choice of model. Influence is often assessed conditionally upon the selected model. However, this does not take the model selection process into account. Nowadays, it is feasible to use an unconditional criterion to determine the influence of each case on the selection procedure. A number of examples are discussed to illustrate the differences between these approaches. Heuristics are developed to explain the examples. We conclude that, although the conditional approach gives valuable information about the selected model, the use of the unconditional approach can lead to greater insight about the influence of individual observations on the process of model selection.

Keywords: Cook's distance, all subsets regression, stepwise regression, regression diagnostics, model selection.

*Supported by NSERC (Canada) and FCAR (Québec)
†Supported by Hatch Grant 151410 NYF

1 Introduction

Variable selection techniques are widely used to determine which variables are "important" predictors, find a reduced set of predictors, or provide better prediction by avoiding overfitting.

Given these goals, it is clearly undesirable for the final model to depend strongly on only a few observations. Measures of influence are therefore very important for model building. In this article we will examine the use of a "leave-one-out" measure of changes in predicted values to assess influence of individual observations in model building.

In most practical variable selection problems, there is some degree of multicollinearity among the independent variables. When the degree of multicollinearity is high, it is well-known that small perturbations of the data induce large fluctuations of the regression coefficients. However, the predicted values from the estimated regression equations may be very stable. For this reason, it appears to be more useful to define model selection in terms of sets of predicted *values*, rather than in terms of the predictor *variables*. When the goal of the investigation is to determine a set of important variables, rather than prediction, all subsets regression can then be used to determine subsets of the predictor variables which produce approximately the same predicted values.

Influence measures for model selection should also, therefore, be measures of how predicted values change with changes in the data. Measures of influence of this type have been developed in the context of ordinary multiple regression where no selection takes place. The idea is that the influence of a case can be determined by leaving that case out of the estimation procedure, and then computing the distance between the predicted values from the full data set, and the reduced data set. (Here a case refers to the values of the variables for a particular experimental unit.) Commonly used measures of this type include Cook's distance (Cook, 1977a) and DFFITS (Belsey, Kuh and Welsch, 1980). A case is declared to be influential if this distance is "large", where size is determined by comparison with some reference value. The differences between these methods are essentially differences in definition of the reference values (Cook and Weisberg, 1982).

On the other hand, there is little explicit advice in the literature on how to assess influence when the model fitted is chosen by a variable selection procedure. One approach is to compute diagnostics (conditionally) on the selected model (for example, Neter, Wasserman, and Kutner, 1985 and Peña and Ruiz-Castillo, 1984). Alternatively, Chatterjee and Hadi (1988) studied the impact of simultaneously omitting a case and a variable from the full model. Weisberg (1981) introduced a statistic for allocating Mallow's C_P, Mallows (1973), to each

case. Weisberg's statistic may be useful in choosing the model least affected by a small subset of cases, from several models with similar values of C_P. However, none of these approaches address directly the model selection aspect of the problem.

Influence measures based on distances between predicted values are readily extended to model selection problems in a manner which accounts for the selection process. In this article we will discuss the use of Cook's distance (defined below) for assessing influence in model selection. However, we expect that the heuristics will apply equally well to other measures based on predicted values.

Ordinary multiple regression is linear in the data. As a result, predicted values for the reduced data sets can be computed by linear updating of the predicted values from the full data set and influence statistics can be computed from the full data set. Model selection, however, is a highly nonlinear procedure. If the selection procedure is taken into account, the predicted values for the reduced data sets cannot be computed by linear updating of full data set predicted values, or of the values predicted for the model selected from the full data set. As a result, influence statistics are not readily computed from the full or selected model. This point has already been noted in studies of influence for selecting transformations of the data (Cook and Wang, 1983), but seems not to have been emphasized for variable selection.

The approach we advocate requires deleting cases one at a time, and reselecting the model, for each case in the data set. Clearly this is very computationally intensive. However, it is not prohibitively expensive in today's computing environment. One data set analyzed in this paper has 94 data points and 19 predictors. Fitting the 94 required "all subsets" regressions, and computing all of the required summary statistics using New S (Becker, Chambers, and Wilks, 1988) on a Sun Sparc station 1+ required only 7 minutes of CPU time.

The paper is organized as follows: In section 2 conditional and unconditional Cook's distance are defined. The next three sections use real data sets to demonstrate some facts about the influence of individual observations in variable selection. Section 3 illustrates the differences between the conditional and unconditional approaches; in particular, unconditionally influential observations are often not influential conditionally. Section 4 illustrates that procedures leading to the same selected model for the full data set can lead to different measures of unconditional influence. Section 5 illustrates that the unconditional approach cannot be replaced by procedures based on influence measures in the full model or a small set of candidate models. We conclude with a discussion in section 6.

2 Influence Measures

In this paper, we consider the use of Cook's distance, in the context of variable selection. Cook's distance was developed to measure the influence of individual points on parameter estimation in the least squares regression problem $y = X\beta + \epsilon$ where y is an n-vector of values of the dependent or response variable, X is a full rank $n \times p$ matrix of independent or predictor variables, β is a p-vector of unknown regression coefficients, and ϵ is an n-vector of independent Gaussian random errors, with mean zero and unknown variance σ^2.

To define Cook's distance for this problem, we need to define, $n + 1$ data sets. The full data set, W, contains all the data for all the cases. For the i^{th} experimental unit, we also define the reduced data set, W_{-i}, which contains the data for every case *except* case i. We also have $n + 1$ vectors of regression estimates and corresponding vectors of predicted values.

In the discussion that follows, the subscript, "$-i$", denotes models and estimates based on W_{-i}. The superscripts, "F", "s", and "(i)" denote the full set of predictor variables, and selected variables based on W and W_{-i} respectively.

We define b^F as the set of estimated regression coefficients computed from W using all the predictor variables, with corresponding predicted values, $\hat{y}^F = Xb^F$. The set of estimated regression coefficients computed from W_{-i} is denoted by b^F_{-i}, with corresponding predicted values, $\hat{y}_{-i} = Xb^F_{-i}$. Note that even though the i^{th} case is not used in estimating b^F_{-i}, $\hat{y}_{-i}$ contains a prediction for that case. Then Cook's distance is defined by

$$D_i = (\hat{y} - \hat{y}_{-i})'(\hat{y} - \hat{y}_{-i})/pMSE_F \tag{2.1}$$

where MSE^F is the regression mean squared error of the regression computed from W. Cook's distance is just the Euclidean distance between the two sets of predicted values, standardized by a measure of variance. D_i is often compared to the 50^{th} percentile of a standard F distribution on p and $n - p$ degrees of freedom, (Cook, 1977a,b), as a heuristic means of determining cases for which it is unusually large. Weisberg (1985) suggests using the value 1, which is the limiting value of the F statistic as n and p become large.

We will now define Cook's distance similarly in a model selection context. For a review of model selection, see Hocking (1976), Thompson (1978a,b), and Miller (1984). Two heuristics are generally used for automatic variable selection: stepwise techniques which limit the number of possible regressions to a relatively small number, and the optimization of a criterion which will, in principle, be done over all subsets. Stepwise techniques are based on sequential tests of regression coefficients (but the "tests" are not generally adjusted for the effects

of variable selection). These tests can also be viewed as tests of the incremental change in predicted values when a single variable is added to or removed from the model. Subset selection techniques based on optimizing an objective function usually minimize the "distance" between the predicted values and the true regression function, $X\beta$. For this reason, we will focus on predicted values defined by a subset of the independent variables, rather than on the selected variables.

Variable selection usually results in selecting a set of size $q^s < p$ variables using W. The corresponding parameter vector will be denoted by β^s, and estimates and predicted values will be denoted b^s and $\hat{y}^s$.

Practitioners are usually advised to assess the goodness of the selected model using the diagnostics from the selected model(s) (for example, Neter, Wasserman, and Kutner, 1985). The influence of an individual case is determined by evaluating its influence on the predicted values of the selected model. We call this the conditional approach, and denote Cook's distance for this method by D_i^c. This requires an estimate of β^s computed from W_{-i} which will be denoted b^s_{-i} and a corresponding set of fitted values, $\hat{y}^s_{-i}$. D_i^c is then computed as

$$D_i^c = (\hat{y}^s - \hat{y}^s_{-i})'(\hat{y}^s - \hat{y}^s_{-i})/q^s MSE^s, \tag{2.2}$$

where MSE^s is the regression mean squared error of the selected regression, computed from W.

We argue, however, that influence should be viewed as a measure of distance between the predicted values computed from the full data set, W, and the predicted values computed from W_{-i}, just as it is in multiple regression. We propose that predicted values for W should be computed from the candidate model, but that those for W_{-i} should be computed by reselecting the model from W_{-i}, using the same variable selection procedure. Ideally, heuristics and expert opinion will augment automatic variable selection techniques, and the exact same heuristics and expertise should be used to select the model based on W and on each W_{-i}. However, except in a highly interactive setting, this will usually be too labor intensive. We propose that automatic procedures, such as minimum C_p, minimum PRESS, Allen (1974) and Geisser and Eddy (1979), or a variant of stepwise regression, can be used as an approximation, to determine the most influential points. These can then be assessed in light of the expertly chosen model.

We call this the unconditional method, and denote the unconditional Cook's distance by D_i^u. In the unconditional approach, the predictor variables are selected using W_{-i}. So, for each i there is a parameter vector, $\beta^{(i)}$, with $q^{(i)}$ elements. The corresponding parameter

estimates are computed only from W_{-i} and are denoted by $b_{-i}^{(i)}$. The fitted values, $\hat{y}_{-i}^{(i)}$, are computed for all the cases in W. In an abuse of notation, $\beta^s = \beta^{(i)}$ will be understood to mean that the 2 selected models contain the same predictors.

We compute D_i^u as

$$D_i^u = (\hat{y}^s - \hat{y}_{-i}^{(i)})'(\hat{y}^s - \hat{y}_{-i}^{(i)})/q^s MSE^{\mathrm{F}}, \tag{2.3}$$

where the estimate of variance used in the denominator is computed from the full model using all cases.

In ordinary regression, the standardization by the denominator fulfills two roles: it provides a common standard to compare the values of D_i across all the cases, and also provides a heuristic for comparing the values to a standard F distribution to assess whether any cases are unusually influential. The use of $q^s MSE^F$ in the denominator, provides a common standard for comparison, and MSE^F is an unbiased estimate of the variance when the full model is correct. When $\beta^s = \beta^{(i)}$, the same points should be declared influential by both the conditional and unconditional approaches. This leads to the use of q^s as the number of parameters to use for comparison. (A good model should have a mean squared error close to MSE^F, so the difference in denominators should not be large.) For the same reason, we advocate the use of the 50^{th} quantile of an F distribution with q^s and $n - q^s$ degrees of freedom as the reference value for D_i^u.

We now study the relationship between D_i^u and D_i^c. Adding and subtracting $\hat{y}_{-i}^s$ to the numerator of (2.3) we obtain

$$D_i^u = \tilde{D}_i^c + F_i + 2C_i, \tag{2.4}$$

where $\tilde{D}_i^c$ is similar to (2.2) except that MSE^{F} replaces MSE^s, $F_i = (\hat{y}_{-i}^s - \hat{y}_{-i}^{(i)})'(\hat{y}_{-i}^s - \hat{y}_{-i}^{(i)})/(q^s MSE^{\mathrm{F}})$, and $C_i = (\hat{y}^s - \hat{y}_{-i}^s)'(\hat{y}_{-i}^s - \hat{y}_{-i}^{(i)})/(q^s MSE^{\mathrm{F}})$. So apart for a small difference in the denominator, $\tilde{D}_i^c$ is the conditional Cook's distance, while F_i can be thought of as a measure of the distance between the models selected from W and from W_{-i}. Note that y_i is not used in predicting $\hat{y}_{-i}^s$ or $\hat{y}_{-i}^{(i)}$. If the models β^s and $\beta^{(i)}$ are nested, then the numerator of F_i is the numerator of the corresponding F-test computed from W_{-i} plus the squared difference of the i^{th} predicted value of the two models. In the cases where W_{-i} leads to a different model than W, we should expect F_i to be large.

Cauchy-Schwarz's inequality implies

$$(\sqrt{\tilde{D}_i^c} - \sqrt{F_i})^2 \leq D_i^u \leq (\sqrt{\tilde{D}_i^c} + \sqrt{F_i})^2. \tag{2.5}$$

Therefore if the i^{th} observation is unconditionally influential, it must either be conditionally influential and/or the models selected from W and W_{-i} must be very different in the sense that their predicted values are relatively far apart.

In our experience, the cross-product term C_i of (2.4) is usually small and positive. Thus, when the models β^s and $\beta^{(i)}$ differ, the unconditional influence is usually larger than the conditional influence. However, despite increased influence, the observation is not necessarily declared influential because of a change of model.

The cross-product can, however, take a large negative value leading to $D_i^c < D_i^u$. The cross-product can be decomposed as:

$$C_i = \frac{\hat{e}_i^s}{(1-h_{ii}^s)\sqrt{MSE_F}} H_i^s \frac{(\hat{y}_{-i}^s - \hat{y}_{-i}^{(i)})}{q^s\sqrt{MSE_F}}, \tag{2.6}$$

where $\hat{e}_i^s$ is the i^{th} residual of the model β^s. The quantities h_{ii}^s and H_i^s are, respectively, the i^{th} diagonal element and i^{th} column of the hat matrix, $H^s = X^s(X^{s\prime}X^s)^{-1}X^{s\prime}$, for model β^s, which has predictor variables X^s. So, for example, if the i^{th} observation has a high leverage in model β^s, then its conditional influence may be large, the first part of (2.6) is likely to be large and the second term is approximately $h_{ii}^s(\hat{y}_{-i;i}^s - \hat{y}_{-i;i}^{(i)})$ which can also be large. The sign of C_i depends on the position of the i^{th} observation with respect to its predicted value in different models. In section 3, we show an example of an observation which is conditionally influential due to a high leverage in that model but which is not unconditionally influential.

3 Conditional Versus Unconditional Cook's Distance - Examples

In this section, we use two data sets to illustrate the following two remarks.

Remark 1 *An observation may be conditionally uninfluential while having a large influence unconditionally.*

Remark 2 *An observation may be conditionally influential while having little influence unconditionally.*

The first data set is the Fuel consumption data of Weisberg (1985). There are 50 observations, one per state, and four predictors. The response is the 1972 fuel consumption in gallons per person. The predictors are Tax, the amount of the tax on a gallon of fuel in cents,

Table 1a	Fuel Data			
Case	D_i^u	$\tilde{D}_i^c$	F_i	$2C_i$
Hawaii	2.40	0.32	1.92	0.16
Wyoming	0.88	0.34	0.52	0.01
Alaska	1.07	0.36	0.65	0.05
Table 1b	Hubbard Brook Forest Data			
Case	D_i^u	$\tilde{D}_i^c$	F_i	$2C_i$
6984	0.57	0.88	0.31	-0.62
Table 1c	Berkeley Data			
Case	D_i^u	$\tilde{D}_i^c$	F_i	$2C_i$
201	0.09	0.091	0.00	0.00
210	2.81	0.001	2.77	0.04
216	3.42	0.001	3.44	-0.02
218	2.71	0.010	2.63	0.08
228	2.72	0.007	2.66	0.05

Table 1: Decomposition of D_i^u.

Dlic, the proportion of the population with a driver's license, Inc, the per capita income in the state, and Road, the total length of roads in the state in thousands of miles.

With minimum C_p as the model selector, the model chosen using all 50 observations is Dlic and Inc. For all but three states, $\beta^s = \beta^{(i)}$. Therefore the conditional and unconditional Cook's distance are identical except for the estimate of variance in the denominator. On the other hand, removing either Wyoming or Alaska modifies the selected model by adding the predictor, Road. Likewise, removing Hawaii adds the predictor, Tax. Consequently, the conditional and unconditional measures of influence are different.

Figure 3.1 shows that no state exerts a large influence in the selected model (Dlic and Inc) as the largest conditional Cook's distance is less than 0.5. We are thus led to believe that no state unduly influences our data analysis. However, Alaska, Hawaii and possibly Wyoming, are unconditionally influential, thus illustrating remark 1. The reason for this can be seen in Table 1a. Both states have relatively high values of F_i.

A fuller analysis of the data explains the effects of the individual states on the analysis. Here, the effect of Hawaii is examined in detail.

According to the C_p criterion, three models were comparable: Dlic and Inc (C_p=2.52), Dlic, Inc, and Road (C_p=3.31), and Tax, Dlic, and Inc (C_p=3.53). No other model had a

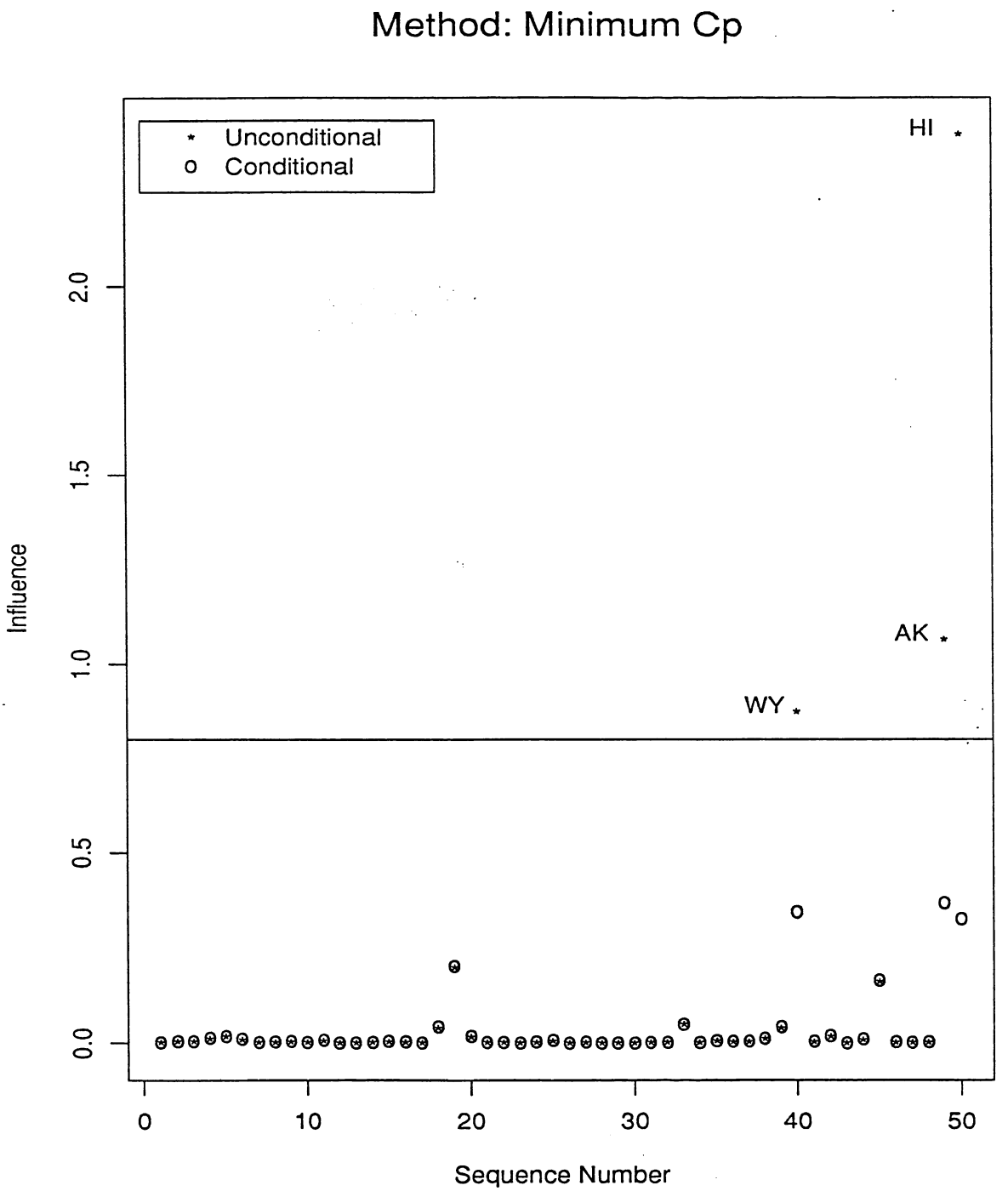

Figure 3.1: Cook's D for the Fuel Consumption Data.

C_p of less than 11. Removing Hawaii changes these statistics considerably: Dlic and Inc (C_p=9.83), Dlic, Inc, and Road (C_p=11.49), and Tax, Dlic, and Inc (C_p=4.33). No other model has a value less than 12. Thus, without Hawaii, the model Tax, Dlic, and Inc is clearly superior to Dlic and Inc. Figure 3.2 shows the added variable plot, (Weisberg, 1985), for the variable Tax in the model Tax, Dlic, and Inc, using W. The least squares line (solid) and the resistant least median of squares line (dashed) of Rousseeuw (1984) are shown on the plot. The difference in slopes indicate that the least squares slope is influenced by one or more unusual points. The location of Hawaii in the lower left corner of the plot, far from the main body of the data, indicates the high influence of this state. Removing Hawaii changes the slope of Tax from -12.95 with a p-value of 0.32 to -31.11 with a p-value less than 0.01.

To illustrate Remark 2, we examine a second data set consisting of 94 observations on 19 predictors. The data are observations of vegetation and site parameters on 47 plots surveyed in Hubbard Brook Experimental Forest in 1971, one year following clear cutting. Two subplots were measured on each plot. (Although the subplots are correlated, this aspect of the data has not been accounted for in this analysis.) The goal of the analysis is to understand how the conditions on the plots following cutting affect the growth of various tree species. In this analysis, the logarithm of the number of pincherry seedlings is regressed on a number of variables, including counts of various herb species (an indication of dampness and local micro-climate of the plot), measurements of the amount of disturbance of the plot by the cutting operation, and location of the plot (including elevation and distance to the nearest edge and western edge of the stand).

Again, we select the model which minimizes the C_p criterion. It contains 8 of the 19 predictors. Three variables, Hayfern, Yviolets, (yellow violets), and Moss, are counts of herbaceous plants, related to site micro-climate. One, Lybirch, (an indicator variable for the presence of yellow birch logs), is related to the type of growth on the site prior to logging. Two, Slash, (the amount of waste tree products left on the site), and Scar, (the amount of scarification of the soil by the logging machinery), are related to the disturbance due to logging. The final two variables, Elev, (elevation), and Edge, (distance to the edge of the forest), are related to the geographical location of the plot.

With 19 highly correlated predictors, it is no surprise that many models have very similar values of the C_p criterion, or that the model selected from W_{-i} is often very different from the model selected from W. In fact for 33 of the 94 plots, $\beta^s \neq \beta^{(i)}$.

Figure 3.3 displays the conditional and unconditional measures of influence. Only plot 6984 is conditionally influential with D_i^c=0.93. Using the conditional approach, we are led to

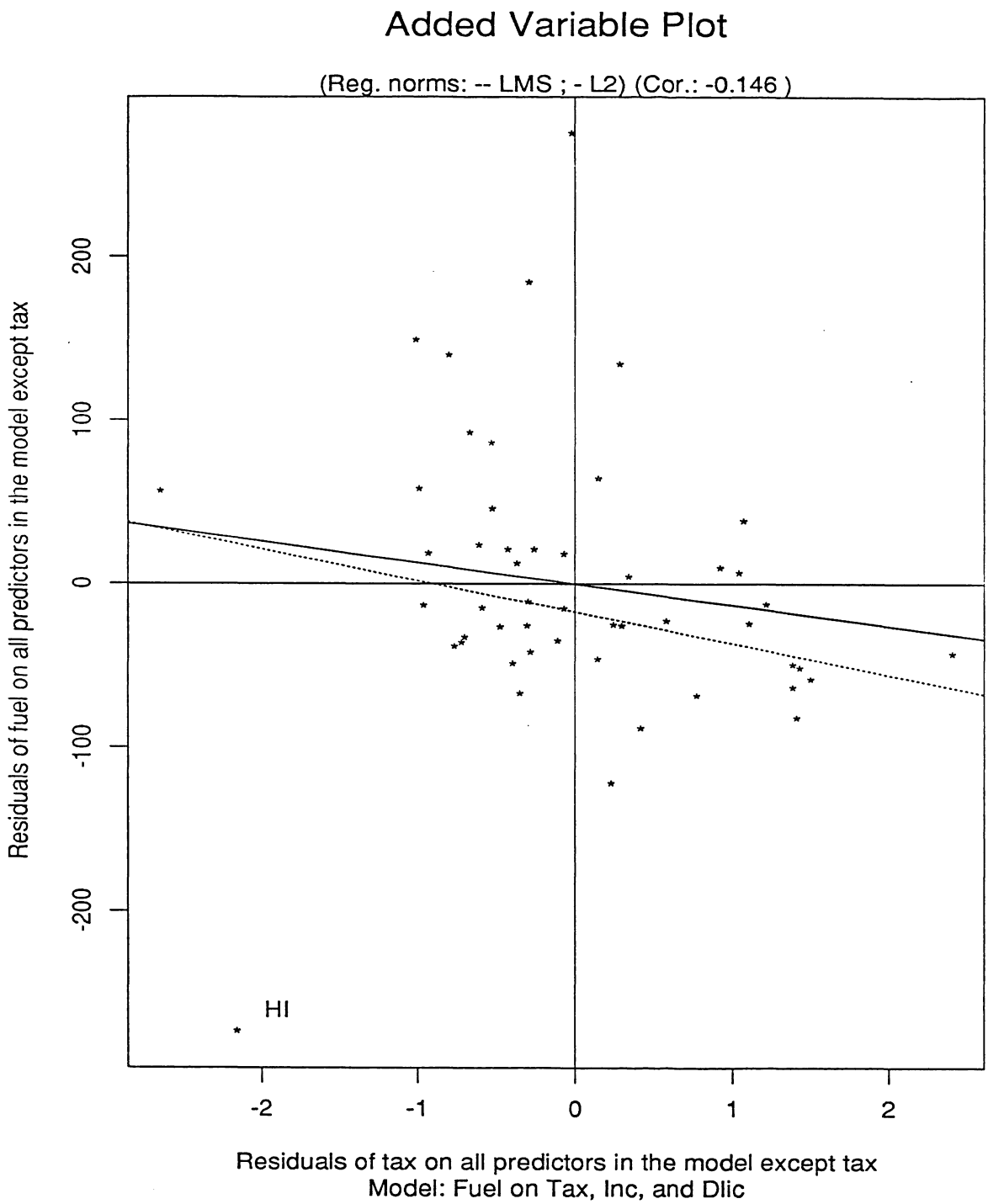

Figure 3.2: Added variable plot of Tax after Dlic and Inc.

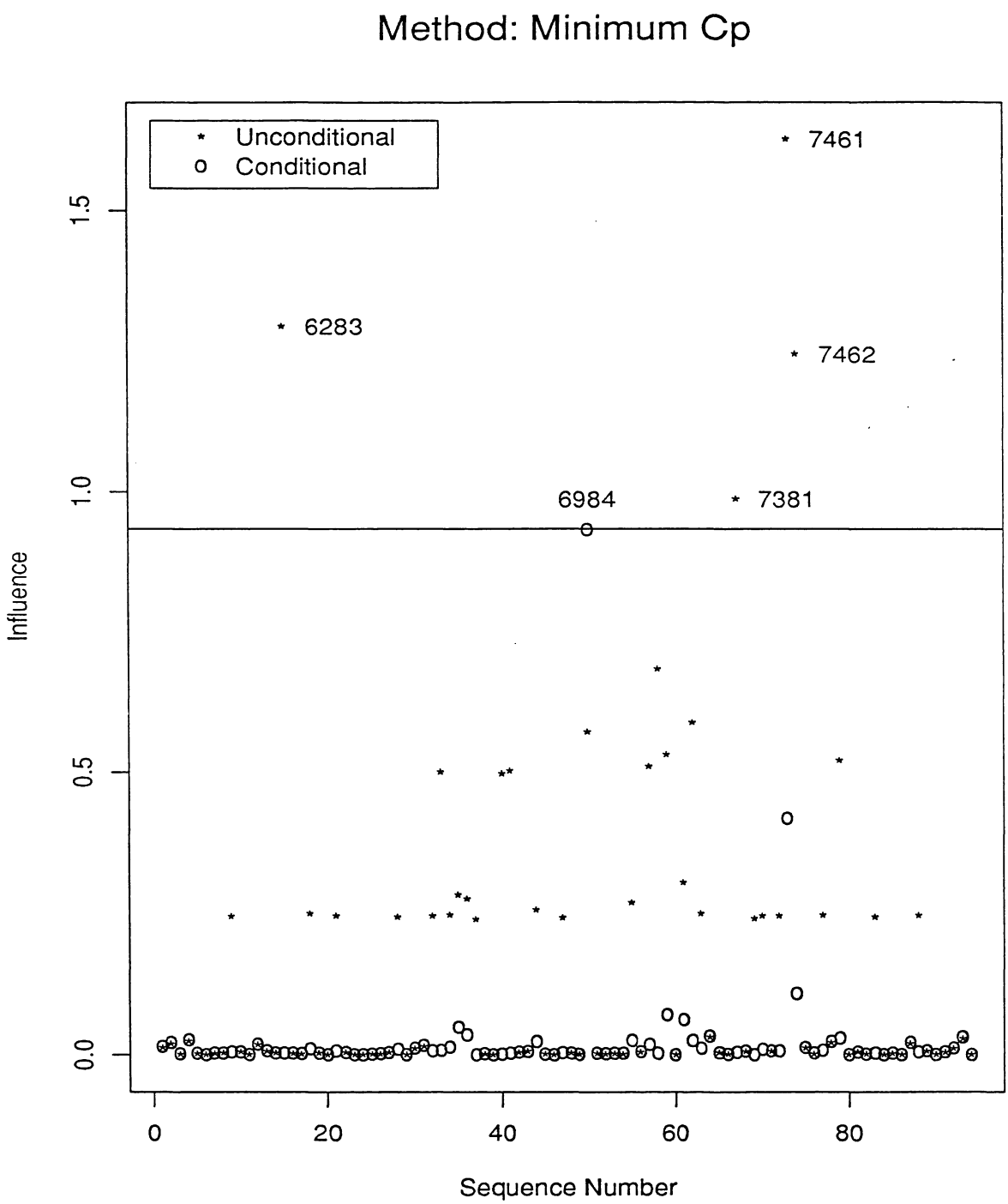

Figure 3.3: Cook's D for the Hubbard Brook Forest Data.

believe that at most one case has a definite influence on our data analysis.

However, the unconditional influence of plot 6984 is less than its conditional influence, and is, in fact, not unusually high.

The decomposition of D_i^u for this case is given in Table 1b. The cross-product term is large and negative. In the authors' experience, this is rare. The magnitude of the term is due to the extremely high leverage of the plot in β^s. This is due to a single variable, Hayfern, which is zero on all but 7 plots. For plot 6984 it is 23, while its next highest value is 6. When Hayfern is included in the model, the removal of plot 6984 has a large effect on the regression coefficients and hence predicted values, reflected by $D_{6984}^C = .93$.

However, when plot 6984 is removed from the data set, the selected model does not include Hayfern. Since Hayfern behaves like an indicator variable for plot 6984, and is not highly correlated with the other predictors, plot 6984 is the only plot affected by the change of model leading to unconditional influence smaller than conditional. Since $\beta^{(6984)}$ is the only selected model which does not include Hayfern, the presence of Hayfern in the selected model appears to be entirely due to this plot. Although Hayfern is selected because of this plot, it does not affect the prediction of the other plots. However, if we are interested in determining the important predictors for this problem, including a variable on the basis of a single plot is undesirable. This shows the value of examining cases for which D_i^c and D_i^u differ.

Figure 3.4 is an added variable plot of the variable Hayfern in the model selected using all the cases. As before, the least squares line (solid) and the least median of squares line (dashed) are indicated. The least squares line is strongly affected by plot 6984 as the large difference in slopes indicates. It is evident that the correlation in the added variable plot is mostly explained by plot 6984. Removing that case changes the slope from -0.06 with a p-value of 0.08 to 0.039 with a p-value of 0.77.

Figure 3.3 also shows that for most cases D_i^u is almost 0. These correspond to the cases for which $\beta^{(i)} = \beta^s$. The 33 other cases all have $D_i^u > 0.24$. Four cases are unconditionally influential: plots 6283, 7381, 7461, and 7462.

4 Unconditional Influence and the Model Selection Procedure

In the previous section, differences between conditional and unconditional measures of influence have been demonstrated for a single variable selection technique, in this case, all subsets regression using the C_p criterion. In this section, we demonstrate that D_i^u also depends on

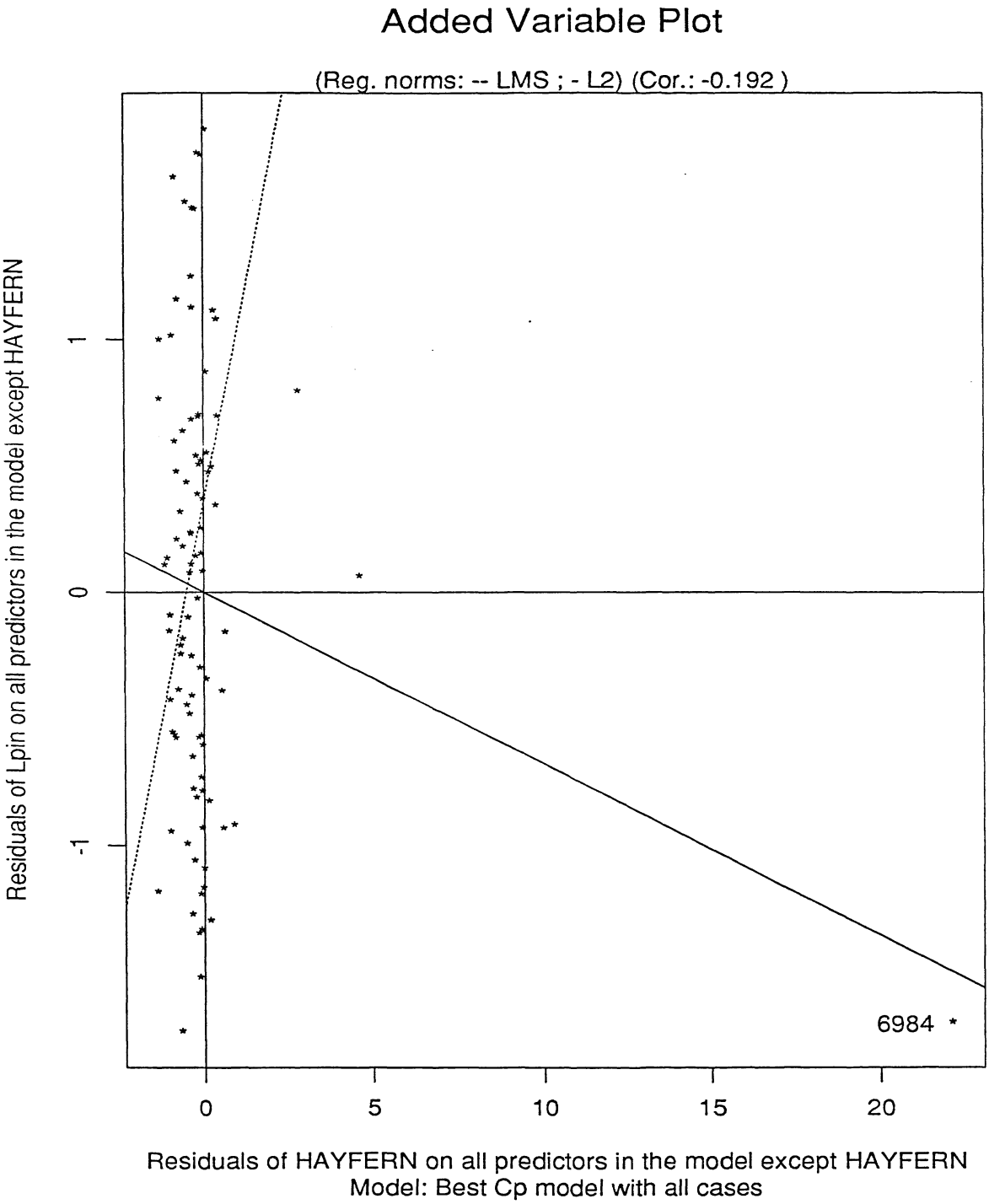

Figure 3.4: Added variable plot of Hayfern in the model selected by all cases.

the variable selection method.

The conditional measure of influence is a function of the data and the selected model, whereas the unconditional measure is a function of the data and the method of selecting a model. So, if two different methods of model selection lead to the same model, the conditional measure will give the same answer. However, the unconditional measures of influence may give different answers, as the models selected from W_{-i} may differ. The example of this section illustrates this point.

Remark 3 *Even if several methods end up with the same model, leading to identical conditional influence measures for all methods, they might have different unconditional influence.*

In the previous section, minimum C_p was used to select a model for the Fuel data. In this section forward selection and backward elimination, are also used. The three methods choose the model with Dlic and Inc from W. Hence, the conditional influence measures are identical for the three models. As we have already seen in figure 3.1, no state is conditionally influential. The statistician might therefore feel comfortable with this model, as the three different methods have selected it and no single case is conditionally influential.

However, we have seen that Wyoming, Alaska and Hawaii are unconditionally influential when the method of selection is minimum C_p. When forward selection or backward elimination is used with an F-to-enter or remove value of 4, only Hawaii is influential. For the 49 other states, forward and backward selection continue to pick the model with Dlic and Inc, and so their forward selection unconditional influence is identical to their conditional influence.

So, Hawaii is unconditionally influential according to the three methods, but Wyoming and Alaska are unconditionally influential only when model selection is done according to the minimum C_p criterion. This can be better understood by examining Alaska in detail. When it is removed, the C_p estimate for the model Dlic and Inc increases from 2.52 to 3.25, whereas that for Dlic, Inc, and Road decreases from 3.31 to 3.08. So the "best" C_p model changes, although the two C_p estimates are now very close. On the other hand, in forward selection, the F values for Dlic, Inc, and Road are 24.4, 10.6, and 1.2 with Alaska and 26.4, 15.7, and 2.2 without it, respectively. So an F-to-enter larger then 2.2, corresponding to a p-value of 0.14 in this case, will not allow the inclusion of the variable Road. In fact, removing Alaska only increases the coefficient of Road from 3.87 with a p-value of 0.27 to 4.92 with a p-value of 0.14. We have seen in the previous section that removing Hawaii had a much more important effect on the selection of a different model according to the C_p criterion. The other two methods of selection have confirmed the importance of adding the variable Tax when Hawaii is removed.

5 Unconditional Influence and Full Model Statistics

An advantage of Cook's distance in multiple linear regression is the fact that it can be computed from statistics based on the model with all observations. It is not necessary to actually compute the predicted values of $n+1$ regressions. In fact,

$$D_i = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right), \tag{5.1}$$

where p is the number of variables in the model (including the intercept if present), r_i is the i^{th} studentized residual and h_{ii} is the i^{th} diagonal element of the hat matrix, $X(X'X)^{-1}X'$. It is therefore very inexpensive to compute. The simple decomposition of D_i is due to the linearity of the estimation procedure. Since D_i^c is Cook's distance of a selected model, it is also simple to compute. On the other hand, model selection is highly nonlinear and so D_i^u cannot be computed using simple updating formulas. Selecting $n+1$ models and computing the corresponding predicted values is the only solution.

Although the cost of computing D_i^u is not prohibitively high in today's computing environment, one may hope to detect all influential observations through less expensive methods that would involve either full model statistics or simple updating formulas. For instance, observations with high leverage and/or large residuals in the full model or in β^s might be likely to have a large unconditional influence. Unfortunately, the example below shows that this will be insufficient to detect all unconditionally influential observations.

Remark 4 *An observation with small leverage and small residual in the full model and in the model selected from W may still have a large unconditional influence.*

We will use a third data set to illustrate this point. It is the data set for boys from the Berkeley Guidance Study found in Weisberg (1985). The response variable is the somatotype, a measure of fatness based on a seven-point scale, at age 18. The predictor variables are weight and height at age 2, 9, and 18, leg circumference and a measure of strength at age 9 and 18. Twenty-six boys took part in the study.

Figure 3.6 contains three points for each subject. On the left y-axis is the scale for the full model Cook's distance (o) and the unconditional Cook's distance based on minimum C_p ($*$). The scale for the full model leverage (X) is on the right y-axis. The model selected from W includes the variables weight at age 2 and 18 and strength at age 18.

Subjects 210, 216, and 228 all have relatively high full model leverage (although none is larger than $2p/n = 0.84$), very large D_i^u, but very small full model Cook's distance. This

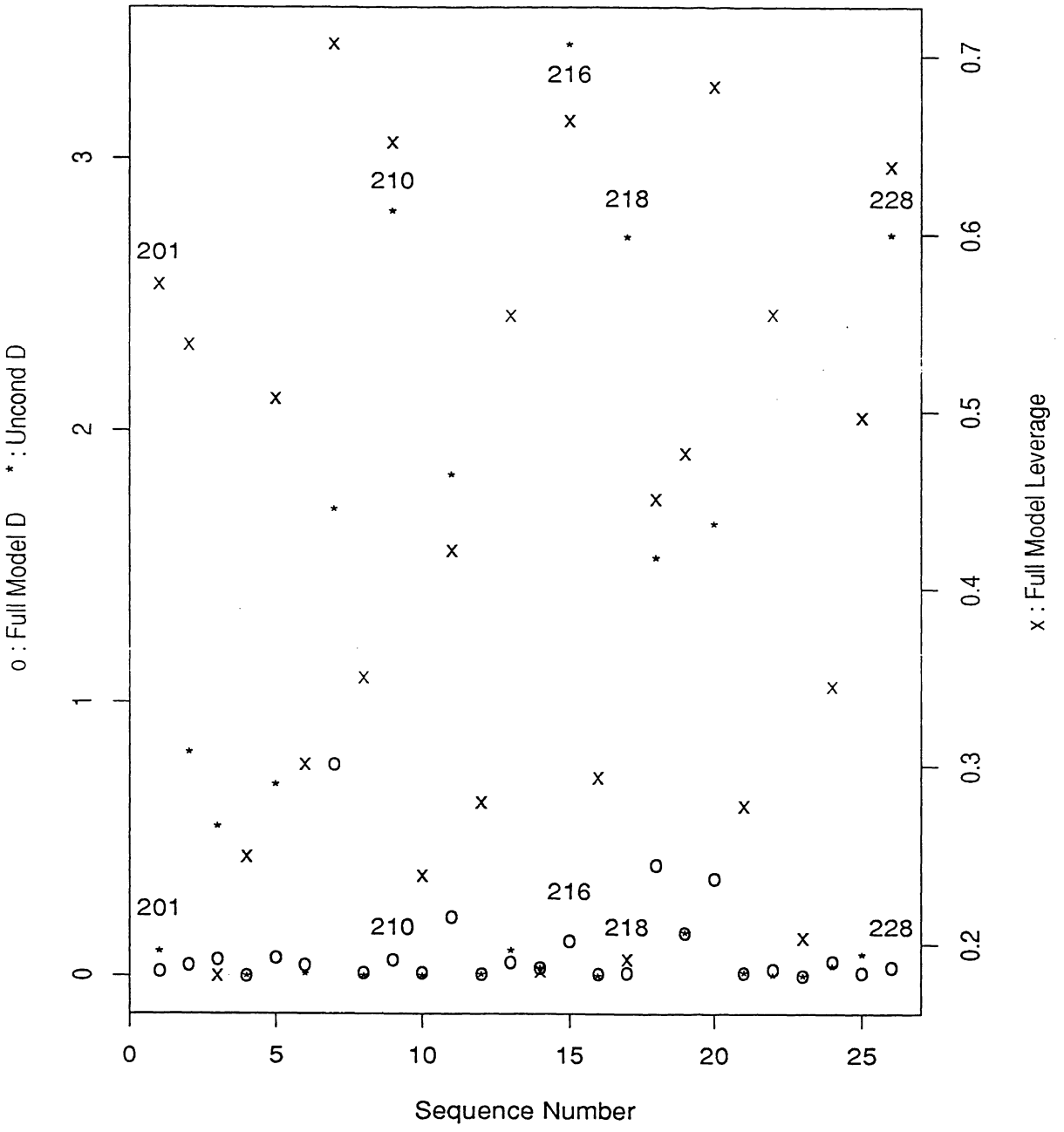

Figure 5.5: Full Model Statistics and Unconditional Cook's D for the Berkeley Guidance Study.

suggests that high full model leverage (although not Cook's Distance) might be related to high D_i^u. But subject 201 shows that it is not always the case. He has relatively high full model leverage, but small D_i^u. This should not be surprising: it is well known that a point can be distant from a high dimensional point cloud, while its projection in a lower dimension is close to the projection of the point cloud.

The major difference between subject 201, and the subjects 210, 216, and 228 is that the model selected from W_{-i} does not change for the first subject, whereas it changes for the last three. Table 1c shows that all 4 subjects have negligible values of D_i^c, but the latter 3 have a large value of F_i.

Not all of the unconditionally influential subjects have high leverage. For example, subject 218 has the third smallest full model leverage (0.19) and his full model studentized residual is -0.72. Likewise, in the model selected from W, he has a leverage of 0.09 and a studentized residual of -0.66. His full model Cook's distance and D_i^c are both 0.01. However, $D_i^u = 2.71$, once again, due to a large value of $F_i = 2.63$.

Equation (5.1) shows a strong relationship between full model leverage and Cook's distance for the full model. However, variables causing full model high leverage may not be important predictors, so that they do not appear in the selected model. Conversely, high leverage in the selected model can be masked in the full model by variables which were not selected. As a result, leverage in the full model need not be associated with high values of D_i^c or D_i^u.

Since large values of D_i^u appear to be associated with a change in the selected model, it would be useful to be able to predict when $\beta^s \neq \beta^{(i)}$. If the model selection procedure consists of choosing the model with the smallest C_p estimate, then one could use simple updating formulas to update the estimate of C_p of the best m models of size k for $k = 1, \ldots, p$ in the hope of finding out the best model. Unfortunately, the number of models to update might be much too large to render this exercise useful.

Remark 5 *The best model(s) after removing case i may differ dramatically from the best model(s) chosen from the full data set.*

To illustrate this remark, consider again the Hubbard Brook Forest data. Figure 3.3 showed that the plots 6283, 7381, 7461, and 7462 are all unconditionally influential. For each number of predictors, we computed the 10 best models according to the C_p criterion. Based on W the best model has 8 predictors. Without plot 6283, the best model also has 8 predictors. The C_p of that model, when computed from W rather than W_{-i}, is not among the 10 smallest for 8 predictor models and there are 36 other models among those computed

that have a smaller C_p. For plot 7381, the model selected from W_{-i} is the 9^{th} best among the models with 7 predictors and 42 other models among those computed have smaller values of C_p. The model selected without 7461 is not among the 10 best of its size and 45 other models are better whereas the model selected without 7462 is the third best of its size and 70 other models among those computed have smaller C_p estimates.

Hence too many models would have to be considered for updates of C_p to be helpful. It seems clear that even though removing an observation and going through the variable selection procedure is expensive to compute, no simple updating formula or simple statistics can be used to alleviate the computational burden.

6 Discussion

In this paper we have introduced two measures of influence for variable selection, conditional and unconditional Cook's distance. We have illustrated why the unconditional approach, while computationally intensive, seems better suited to the goals of variable selection procedures. However, conditional diagnostics are useful in investigating the selected models.

It is important to notice that, although D_i^u is not readily computed from residuals and leverages based on the full or selected model alone, once a point has been identified as unconditionally influential, the reasons for this influence can be determined using conventional conditional diagnostic tools such as leverage and added variable plots. Likewise if D_i^c is much larger than D_i^u, this case is worthy of further investigation. As in ordinary multiple regression, the practitioner can then determine the desirability of using influential cases in the analysis. However, the diagnostic tools will be used on both models β^s and $\beta^{(i)}$ when these differ. Compared to the use of conditional measures, which examine only β^s, this provides the practitioner with a fuller picture of how the final model is affected by which cases are included in the model, and should lead to more informed model selection.

Other measures of change in predicted values when a single point is left out of the data set exist. For instance, DFFITS of Belsey, Kuh, and Welsch (1980) is defined by

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{-i;i}}{s(i)\sqrt{h_{ii}}}, \tag{6.1}$$

where $\hat{y}_{-i;i}$ is the i^{th} predicted value computed without case i, $s(i)$ is the variance estimate of the model computed from W_{-i} and h_{ii}, the leverage, is the i^{th} diagonal element of the hat matrix $X(X'X)^{-1}X'$. Unlike the denominator for D_i, the reference denominator for DFFITS_i varies with i. However, for ordinary regression, the distribution of DFFITS_i, unlike that of

D_i, can readily be computed. While a similar measure can readily be defined for influence in variable selection, the distribution is no more tractable than that of D_i. Therefore, D_i, with its common denominator, seems more useful.

These measures have also been generalized to assess the influence of a group of points on the predicted values in ordinary multiple linear regression (see, for instance, Cook and Weisberg, 1982). Unconditional Cook's distance can easily be generalized to the multiple removal of points, although, as in ordinary multiple regression, it rapidly becomes computationally expensive to consider all possible groups.

Variable selection in regression is one of the most used statistical techniques. Although the estimation aspect of that technique has been studied extensively, inference and assessment of influence has always been done conditionally on the selected model for lack of proper techniques that incorporate the selection part of the problem. As is shown here, assessment of influence can be done satisfactorily and use of unconditional Cook's distance helps in understanding the data, and in the choice of model.

Another approach to influence is the use of robust estimation techniques. Instead of identifying potentially influential cases, least squares regression is replaced by high breakdown methods, such as the Least Median of Squares of Rousseeuw (1984). This way, no single observation can exert too much influence, so that influence diagnostics are unnecessary. One the other hand, influence diagnostics can give important information about unusual, and possibly scientifically significant, cases.

Robust techniques are computationally intensive. To our knowledge, they have not, for this reason, been applied to model selection problems. It is likely that computation of unconditional influence diagnostics using least squares fitting will be less computationally intensive than use of robust regression in model selection. Also, the breakdown properties of robust regression, in the model selection setting, are currently unknown. Therefore, in the short term, we expect that unconditional influence measures will be more useful. In the longer term, robust variable selection and influence diagnostics will undoubtedly both be useful tools in the data analyst's toolkit.

References

Allen, D. M. (1974) The Relationship Between Variable Selection and Prediction. *Technometrics,* **16** 125–127.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988) *The New S Language: A Program-*

ming Environment for Data Analysis and Graphics, Pacific Grove: Wadsworth.

Belsey, D. A., Kuh, E., and Welsch, R. E. (1980) *Regression Diagnostics*, New York: Wiley.

Chatterjee, S. and Hadi, A. S. (1988) Impact of Simultaneous Omission of a Variable and an Observation on a Linear Regression Equation, *Computational Statistics & Data Analysis*, **6** 129–144.

Cook, R. D. (1977a) Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15–18.

- (1977b), Letter to the editor. *Technometrics*, **19** 349.

- and Wang, P. C. (1983) Transformations and Influential Cases in Regression. *Technometrics*, **25** 337–343.

- and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.

Geisser, S., and Eddy, W. F. (1979) A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, **74** 153–160.

Hocking, R. R. (1976) The Analysis and Selection of Variables in Linear Regression. *Biometrics*, **32** 1–49.

Mallows, C. L. (1973) Some Comments on C_p. *Technometrics*, **15** 661–676.

Miller, A. J. (1984) Selection of Subsets of Regression Variables (with discussion). *Journal of the Royal Statistical Society series A*, **147** 389–425.

Neter, J., Wasserman, W., and Kutner, M. H. (1985) *Applied Linear Statistical Models*. (2nd ed.), Homewood: Irwin.

Peña, D., and Ruiz-Castillo, J. (1984) Robust Methods of Building Regression Models—An Application to the Housing Sector. *Journal of Business & Economic Statistics*, **2** 10–20.

Rousseeuw, P. W. (1984) Least Median of Squares Regression. *Journal of the American Statistical Association*, **79** 871–880.

Thompson M. L. (1978a) Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. *International Statistical Review*, **46** 1–19.

Thompson M. L. (1978b) Selection of Variables in Multiple Regression: Part II. Chosen Procedures, Computations and Examples. *International Statistical Review*, **46** 129–146.

Weisberg, S. (1981) A Statistic for Allocating C_P to Individual Cases *Technometrics*, **23** 27–31.

Weisberg, S. (1985) *Applied Linear Regression* (2nd ed.), New York: Wiley.